e-CRM, web semantic propensity models and micro-data-mining : an application of Kernel Discriminant Analysis to the Glam on Web case

Furio Camillo – Caterina Liberati

Dipartimento di Scienze Statistiche - Alma Mater Studiorum, Università di Bologna - Italy

fcamillo@stat.unibo.it – liberati@stat.unibo.it

Federico Neri[1]

Synthema s.r.l. – Pisa – Italy

federico.neri@synthema.it

1. Introduction

The change of philosophy in the strategic management of the company, the infusion of technology in the management of the information invaded all the business processes, with consequent organizational modifications oriented to the formation of one structure made to develop knowledge, from the inside to the outside of the enterprise. This process (called information processing) allows simultaneous use of the information in different business and extra-business areas and moreover the representation and use of the data in different ways thanks to the internet use that now is considered one of the most important channel for the transmission and the sharing information.

The role and the importance of the Customer Relationship Management (CRM) in this contest today is unquestionable.

In particular this is true when we consider the e-commerce. This is a particular market in which the businesses can keep in touch with their own costumers very easily because customer service costs are very chip and the computer solutions allow to monitor the user in real time to offer them products or services that are more close to their needs through an affective and personalized communication. Hence for the web world is developed the concept of e-CRM (electronic Customer Relationship Management) that is the electronic management of the relationship with the customer. e-CRM is the business ability to take care of its own clients through internet and tools purposely designed for the on-line world.

In this scenario data mining tools for classification pattern are necessary to find out the most profitable segments of the customer database. Moreover, because of the feedback and the dialogue with the web customers is more simple to realize the business efforts to extract the pattern clickstream data are done to make predictions based on the past behaviour, for building delve customer profile that guide the one–to–one marketing strategies.

[1] Federico Neri is the co-author which processed, specifically, textual data using the tools copyrighted by Synthema s.r.l.

This work starts from the analysis of micro-data saved in a Customer Data Base (CDB) in which it's possible to read not only the personal information but the values and all characteristics of the past purchases : it's known that this is dramatically important to develop a strategy to management customer-business relationship. Naturally for the e-commerce market collecting information is more simple that in other context because via registration on-line business can get information about customers or potential ones. This information is the fundamental part of the marketing study to make and introduce an on-line catalogue where the latter is conceived as a collection of pictures accompanied by technological and commercial texts written by the company. In our case study we analyse a web catalogue of a fashion company realized as a garments selection described by texts.

The innovative idea of this work is composed of two parts : 1) use of texts as a proxi of internaute path that we think is directly linked to the purchase behavior : 2) the use of a new robust algorithm of classification Kernel Discriminant Analysis (KDA) for getting a correct assignment of (new) the subjects to the group to which they belong. In particular for this latter aspect we make a comparison between standard classification technique as Fisher Discriminant Analysis (DA) and KDA showing how use of non-linear transformation of the input variables facilities the building and the identification of the grouping structures especially when the classes surfaces are overlapping.

This paper is born thanks to a young company that decided to experiment new data mining methodologies and to learn from the data the new tendencies.

"Glam On Web" in fact is a company drawing on a wealth of diverse multimedia expertise it is also a multimedia project agency, laying particular emphasis on the integration of communication and technology.

The interaction between statistic research and new contest of application as e-commerce could make this paper very interesting.

2. The case study

Our case study belongs to the e-commerce world : we analysed data and texts coming from Glam on Web portal to discover if customers (buyers) and visitors of the site (no buyers) have the same behaviour. The context of the analysis is the click-stream that is the process of collecting, analyzing, and reporting aggregate data about navigation done by a user in the web site.

As we said before our target variable is the purchase made by the visitors that we can code with a binary string of values Y=[1,1,1,0,0……,0,1,1] with 1=buy and 0=no-buy : where the explicative variables of this models could be information coming from Customer Data Base (CDB) that is related to social-demographic data, user path (succession of mouse clicks that each visitor makes), structural indexes of surf and in the case of customers their historical purchase behaviors, together with their semantic basket (texts clicked and/or concepts "married").

This kind of analysis for companies that sell their products on-line is fundamental because allow them to recognize which users visit the site and at the same time give them the possibility to develop new marketing strategies oriented to encourage users to purchase. Generally to collect information about visitors the webmaster of the portal use to push the users to log in to the web site : in this way it can monitor the site and at the same time it can recognize different users when they come to surf on the portal.

3. The Glam on Web portal[2]

The Glam on Web is projected to be a great emotional value site. It is composed by artistic pictures that "describe" products (or the product-lines) in materials and in features as a catalogue and, at the same time, express the concepts and the creativity that the company wants to use for its marketing campaign. This double function is "translated" in a site in which is very easy to surf freely (without any request of log in). This produces a total lack of feedback information about users hence the subjects of our analysis will be the sessions[3] instead of the users.

In fact this strategic choice makes CDB really poor of data so in our model we don't have predictors coming from customer database.

The navigation in the site of Glam on Web occurs as when one flips through a fashionable magazine composed by images and products descriptions.
Each image of the catalogue could be adapted to the user frame according to the target of the internaut. The system dynamically chooses the images (fig. 1).

The showed garments and/or swimsuits photos are equipped by texts which are provided directly from the internal agency who realizes the communication campaigns. The advertising agency furthermore collaborates to the creative garment realization and therefore texts are the emotional and symbolic garment description.

Hence the web site can be used as a virtual visual stimulus system that allows the company to reach its main aims that are selling and gaining customers fidelity.

In the light of this considerations data building process has been more complicated than we expected : web portal frame and the lack of social demographic information required the creation of new variables as structural indexes of users navigation, factors scores coming from a lexical correspondence as a proxi of information relative to the objects clicked (e-Semantic Basket).

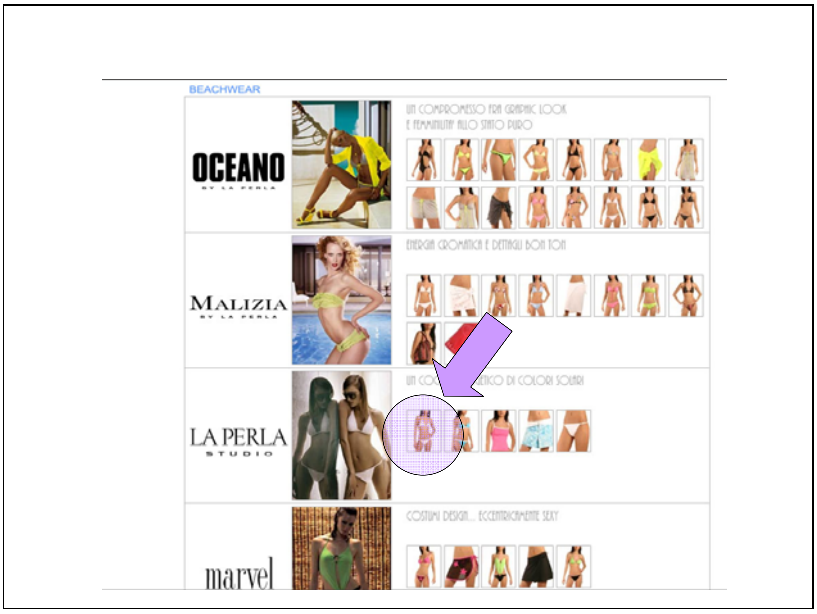

BEACHWEAR
OCEANO
BY LA PERLA
UN COMPROMESSO FRA GRAPHIC LOOK
E FEMMINILITA' ALLO STATO PURO
MALIZIA
BY LA PERLA
ENERGIA CROMATICA E DETTAGLI BON TON
LA PERLA
STUDIO
UN CO… ETICO DI COLORI SOLARI
marvel
COSTUMI DESIGN... ECCENTRICAMENTE SEXY

Fig. 1 *– The web site catalogue : a collection of objects*

[2] The application was supervised by Federico Venturoli (General Manager of Glam on Web)

[3] The session of activity for one user on a Web site during a fixed time-frame

4. Data modeling strategy

Data modeling has been realized in two different ways according to the portal frame the information in hand and the goal of the analysis.

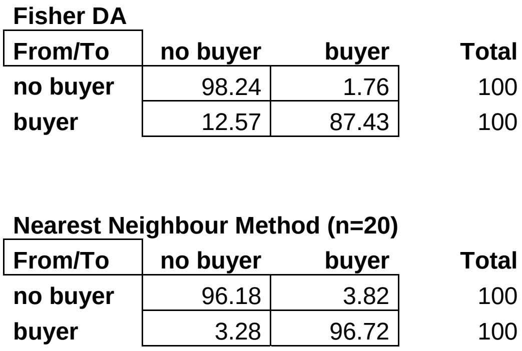
A first specification has been realized using session navigation indexes and the objects classification made by the company, about the web-pages created in the observation period (two weeks of April 2005).

Model I

Buy/no buy (1/0) = f (Navigation variables, Company classification of objects)

The first variables are synthetic measures (mean and standard deviation) that summarize the differences among the sessions in quantitative terms, the others are factors coming from a correspondence analysis of a frequency matrix sessions (rows) objects clicked grouped in macro categories[4] by the company (columns).

The specification of *Model I* belongs to a classical problem of Discriminant Analysis (DA).



Fisher DA

From/To	no buyer	buyer	Total
no buyer	98.24	1.76	100
buyer	12.57	87.43	100

Nearest Neighbour Method (n=20)

From/To	no buyer	buyer	Total
no buyer	96.18	3.82	100
buyer	3.28	96.72	100

Fig. 2 – *Model I : confusion matrixes*

A good classification is obtained using both a classic parametric approach (Fisher DA) and a non-parametric approach (NNM) (Fig. 2). This result derives from a good explicative power of the predictors : the users behavior on the web-site.

Model II

We have to point out some aspects of the *Model I* specification. It is known that one of the limit of the click stream analysis consists in the impossibility for the researcher to estimate a model that can “have life” for a long time. A web site generally is changed many times in a month so it has no sense to use as a predictors only the showed objects. Another weak point of this model is the use of the navigation index coming from the session behavior : it is not possible to make a prediction with this specification we can describe only the status observed. Hence for trying to have a more lasting model we realized a second specification in which we use only the semantic basket as a set of predictors.

[4] Macro categories here mean an objects classification into groups a priori defined according to the technological and visual arguments of the web site.

Buy/no buy (1/0) = f (Semantic Basket)

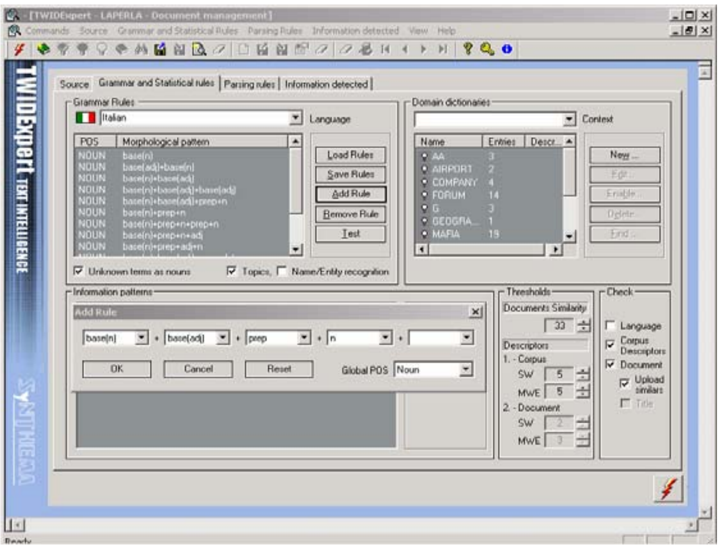
As described above the web site can be used as a virtual visual stimulus system that allows the company to reach its main aims that are selling and gaining customers fidelity.

The textual descriptions of the 402 garments photos was submitted to a structural lexical analysis. The textual space is a sort of virtual space in which the internaut surfed. We think that users with a similar purchase-propensity have been submitted to the same emotional visual (therefore textual) stimuli.

Following this construction we group the 402 garment photos into 7 classes according to a textual clustering process.

The automatic Linguistic Analysis approach used is based on Morpho-Syntactic and Statistical rules. This phase is intended to identify only the significant expressions from the whole raw text. This analysis recognises as relevant terminology only those terms or phrases that comply with a set of pre-defined morpho-syntactic patterns (i.e. : noun+noun and noun+preposition+noun sequences) and whose frequency exceeds a threshold of significance. In fact, a specific algorythm associates an Information Quotient to each detected term, ranking it on its importance. The Information Quotient is calculated taking in account the term, its Part Of Speech tag, its relative and absolute frequency, its distribution on documents. This morpho-syntactic analysis detects significant Simple Word Terms (SWT) and Multi Word Terms (MWT), annotating their headwords, their relative and absolute positions (fig. 3). The detected terms and phrases are then extracted, reduced to their Part Of Speech tagged base form (Raffaelli, 1992) automatically referenced to their hyperonyms, then used as descriptors for documents (Elia e Vietri, 2000).

Indexation based on terminology detection is extremely reliable for managing any type of documentation, expecially if it is technical and scientific. In fact, unfortunately, few of us have complete knowledge about the world. And, in the consequence of this, the meanings we ascribe to words may differ from those ascribed by others. The same happens with lexical tools capable of syntactic parsing, which have always a limited capability of semantic interpretation and disambiguation, if applied to generic corpora. In such situations, these tools cannot pick out the exact interpretation for all expressions in the language. Besides, main terminology - mostly compound nouns – helps "understand" the topic, being intrinsically linked to semantics.



Language: Italian

POS	Morphological pattern
NOUN	base(n)
NOUN	base(adj)+base(n)
NOUN	base(n)+base(adj)
NOUN	base(n)+base(adj)+base(adj)
NOUN	base(n)+base(adj)+prep+n
NOUN	base(n)+prep+n
NOUN	base(n)+prep+n+prep+n
NOUN	base(n)+prep+n+adj
NOUN	base(n)+prep+adj+n

☑ Unknown terms as nouns ☑ Topics, ☐ Name/Entity recognition

Name	Entries	Descr...
AA	3	
AIRPORT	2	
COMPANY	4	
FORUM	14	
G	3	
GEOGRA...	1	
MAFIA	19	

base(n) + base(adj) + prep + n +

Global POS: Noun

Documents Similarity: 33

Descriptors
1. - Corpus
SW 5
MWE 5
2. - Document
SW 2
MWE 3

☐ Language
☑ Corpus Descriptors
☑ Document
☑ Upload similars
☐ Title

Fig. 3 – *TwidExpert : the used tool.*

The classification is made by TEMIS Insight Discoverer Clusterer, according to the K-Means approach. The application dynamically discovers the thematic groups that best describe the analysed documents by the Unsupervised Clustering approach.

In our case, the process performed on more than 400 lingerie descriptions. The Text Mining analysis automatically detects seven groups of documents, each with highly distinctive characteristics (fig. 4). The results shown in are similar to what everyone would expect from reading these type of documents : the group of documents which deals with brassières are grouped or strictly linked to strapless or push-up bra documents, whilst suspender belts are linked to stockings.

Cluster 1	tulle ; poliestere ; stretch ; reggiseno ; viscosa ; poliammide ; perizoma ; spallina ; elastam ; dettaglio_di_stile ; raso ; coppa ; short ; stile ; cotone ;
Cluster 2	bikini ; lycra ; bikini_a_triangolo ; costume ; triangolo_in_lycra ; triangolo ; pareo ; lycra_opaco ; bikini_a_triangolo_in_lycra ; ferro ; costume_intero ; fluo ; ferro_in_lycra ; lycra_lucido ; coppa ;
Cluster 3	pizzo ; leavers ; motivo ; balza ; gusto ; prezioso ; retrò ; riga ; calza ; seta ; leavers_dal_gusto ; prezioso_pizzo ; motivo_di_righe ; balza_alternata ; raffinato_motivo ;
Cluster 4	seta ; pura_seta ; notte ; show ; perla ; fragranza ; regalo ; pastello ; femminile ; corsetto ; cd ; volume ; corpo ; celeste ; sfumatura ;
Cluster 5	puro_cotone ; capo ; capo_della_collezione ; collezione ; manica ; grigiosport ; sport ; tempo_libero ; t-shirt ; t-shirt_in_puro_cotone ; tempo ; libero ; collo ; manica_lunga ; cotone ;
Cluster 6	elastan ; cotone ; boxer ; slip ; jersey ; maglia ; elastico ; canotta ; vita ; elastico_in_vita ; perla ; jersey_di_cotone ; stretch ; spallina_elastica_regolabile ; perizoma_in_cotone ;
Cluster 7	cucitura ; contrasto ; collant ; cucitura_a_contrasto ; den ; tasca ; tassello ; vita ; grigiosport ; banda ; boardshort ; collant_elasticizzato ; comfort ; pigiama ; informal ;

Fig. 4 – *Significant characteristic headwords of every cluster*

The last step for building the exogenous variables of the *Model II* consists in the extraction of factor scores for each internet-user calculated on the lexical matrix *(users)*(1-7 cluster of objects)*[5].

The factors scores coming from a lexical correspondence analysis (LCA) as a proxi of information relative to the clicked objects and of the semantic content of the e-trip of the user. Therefore the *Model II* specification become :

Buy/no buy (1/0) = f (Semantic Basket=factor scores of the LCA)

[5] The *ij-th* element of this matrix is the frequency of visit, for each *i-th* web-session, of a textual object of the *j-th* cluster of objects. This frequencies are weighted with a relative index of characterization of each cluster by the specific words.

Fisher DA

From/To	**no buyer**	**buyer**	**Total**
no buyer	88.50	11.50	100
buyer	57.69	42.31	100

Nearest Neighbour Method (n=20)

From/To	**no buyer**	**buyer**	**Total**
no buyer	81.10	18.90	100
buyer	30.77	69.23	100

Fig. 5 – *Model II : Confusion matrices*

Observing the confusion matrixes showed above (fig. 5) it's clear this latter model has less classification power respect to the first one. That is due to the exclusion in *Model II* of all behavior variables : hence the model loses a little bit in description capacity but gains in prediction estimation. In fact the textual factors represent a more stable reference space in which is always possible to add new objects without changing the set of predictors[6].

3. Using the Kernel Discriminant Analysis

In this paper we propose a data mining tool for supervised classification pattern which is a nonlinear extension of LDA based on kernel functions : Kernel Discriminant Analysis (KDA). The main idea of the kernel method is that without knowing the nonlinear feature mapping or the mapped feature space explicitly, we can work on the feature space through kernel functions, as long as the problem formulation depends only on the inner products between data points. This is based on the fact that for any kernel function κ satisfying Mercer's condition (Cristianini, Shawe-Taylor, 2000) there exists a mapping Φ such that

$$<\Phi(a),\Phi(b)>=\kappa(a,b)$$

where <,> is an inner product in the feature space transformed by Φ (Burges, 1998)). The improvement performing LDA in the feature space instead of the original input space is that it allows us to obtain a decision functions which are nonlinear in the input space but that become linear in the feature space.

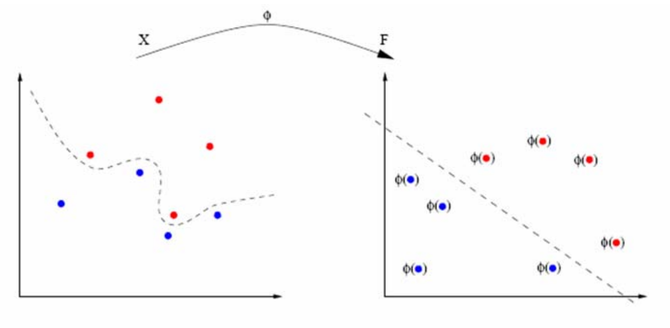


Fig. 6 – *Decision boundary in the Input Space vs. Feature Space*

[6] A new object is projected in the space found with the lexical correspondence computing its coordinate respect to such space or giving it the same coordinate of an other object from whose it is more close.

The reformulation of DA in the feature space is very similar to the LDA case the decision boundary is obtained maximizing the ratio :

$$J(w) = \frac{w' S_B^{\Phi} w}{w' S_W^{\Phi} w}$$

where S_B^{Φ} S_W^{Φ} are Between and Within covariance matrixes in the feature space.

The use of KDA presents some subjective choices which are still open problems : the Kernel Function choice and its parameters.

In this paper we propose the use of Information Complexity theory (Bozdogan, 2000). In particular we referred to the index of complexity (ICOMP) as tool for the model selection in case of KDA (Bozdogan, Camillo, Liberati, 2005).

$$ICOMP\left(\hat{\Sigma}_W\right) = np \log 2\pi + n \log\left|\hat{\Sigma}_W\right| + np + 2C_{1F}\,(\hat{\Sigma}_W)$$

where n is number of objects, p is number of variables, C_{1F} is the measure of complexity of a covariance matrix $\Sigma_W = S^{\Phi}/n$ based on Frobenius norm.
Testing 150 models (50 for each different Kernel function : polynomial, RBF, Cauchy) we founded the best solution in the Cauchy kernel (width = 0.001).

KDA hibridizated with NNM (n=20)

From/To	no buyer	buyer	Total
no buyer	97.00	3.00	100
buyer	26.92	73.08	100

***Fig. 7 –** Model II using KDA (Cauchy Kernel) : confusion matrixes*

The improvements gained applying the KDA on *Model II* are evident. KDA analysis is be able to overcome the limitations of a linear approach that in real case, as this one, are often important. Moreover choosing to adjust the discriminant scores obtained with a Nearest Neighbor process (with n=20) we got the best results showed above (Fig. 7).

5. Problems and future developments

The strategy that we showed is really powerful and interesting but it presents more than one problem to be applied that we can’t ignore :

- *Technological problem* : today the companies that work in e-commerce need solutions (just in time) that are easy and fast to score the new customers as Fisher solution is : KDA is far from to be an automatic/semiautomatic process.
- *Sampling problem* : the context of this kind of analysis is the prediction of a rare event : in general, in the e-commerce business, we have a lot of web-sessions which 0.4%-0.15% buyers and the others only visitors.
- *Generalizing the textual classification of the web-objects :* the allocation of a new object described by a different texts and concepts not expressed in the clusters found.

So we are working at the same time on :

- Testing this new data mining techniques in a operative context evaluating the practicable solutions in a real data analysis system.
- In e-CRM (in our case study) could be useful to use semiometric approach[7] of text mining for codifying and analyzing visual web stimuli, according to a general socio-psychological landscape for the "words" interpretation (Lebart, Piron, Steiner, 2003).

References

Baudat G. and Anouar F. (2000). *Generalized discriminant analysis using a kernel approach*. Neural Computation.

Bozdogan H. (2000). Akaike's Information Criterion and Recent Developments in Information Complexity. *Journal of mathematical and Psychology*.

Bozdogan H., Camillo F., Liberati C. (2005). On the Choice of the Kernel Function in Kernel Discriminant Analysis Using Information Complexity. submitted to *Proceeding Cladag2005,* Parma.

Burges C. (1998). *A tutorial on support vector machines for pattern recognition*. Data Mining and Knowledge Discovery.

Camillo F. (2004). Clickstream analysis, Semiotic Interpretation and Semantic Text Mining for a distance measurement on the hipertextual map of an Internet portal. In SPIROS SIRMAKESSIS. *Text Mining and its applications*, ISBN : 3-540-20238-2, HEIDELBERG, Springer, Germany.

Camillo F. (2005). *Problematiche di gestione del rapporto con gli utenti di un portale web in un'ottica di CRM*. Statistica applicata, ISSN : 1125-1964.

Cristianini N., Shawe-Taylor J. (2000). *Support Vector machines*. Cambridge University Press.

Elia A., Vietri S. (2000). Electronic dictionaries and linguistic analysis of Italian large corpora. *JADT 2000, 5th International Conference on the Statistical Analysis of Textual Data*, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland : 2-4.

Lebart L., Salem A., Berry A. (1998). *Exploring Textual Dat*a. Kluer Academic Publishers, Dordrecht Boston, London.

Lebart L., Piron M., Steiner J.F. (2003). *La Semiometrie*, Dounod Parigi.

Mercer J. (1909). *Function of positive and negative type and their connection with the theory of integral equations*. Philosophical Transactions Royal Society London.

Raffaelli R. (1992). *An inverse parallel parser using multi-layerd grammars*. IBM Technical Disclosure Bullettin, 2Q.

Vapnik V. N. (1995). *The Nature of Statistical Learning Theory*. Springer.

http://www.glamonweb.it

[7] *Semiometrie* is a list of 210 words properly declined which allows the reconstruction of psycho-cultural models that constitute the subconscious system of choice and identification of desires of European citizens. In this sense it is based on the principle that words are not only significant of things but they refer to values and affections to which a single or a group of people are related